Web mining

Wstęp

Założenia dla systemu

Zbudować system, który upraszcza czytanie wpisów na blogach/forach itp. System powinien pobierać wpisy z określonego forum a następnie znaleźć *k* najbardziej reprezentatywnych wpisów i przedstawić je użytkownikowi.

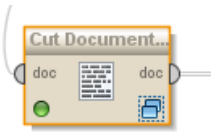
Do realizacji zadania niezbędne są następujące kroki:

1. Wczytanie strony zawierającej forum
2. Podział strony z forum na niezależne bloki, z których każdy będzie reprezentował pojedynczy wpis na forum
3. Zamiana wpisu na reprezentację typu „worek słów" (ang. bag of words)
4. Filtracja wpisów – usunięcie błędnych/pustych wpisów
5. Wyszukiwanie wpisów podobnych – klasteryzacja dokumentów
6. Wyodrębnienie dokumentów wzorcowych
7. Wyświetlenie wyników

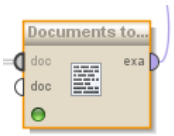
Opis operatorów lub ich grup

Operatory związane z przetwarzanie tekstów / pobieraniem dokumentów z sieci

wczytanie strony. Jako parametr należy podać stronę do wczytania. Uwaga wymaga dodatku Web Mining



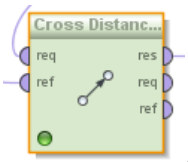
dzieli dokument na fragmenty, tworząc z każdego z fragmentów osobny dokument. Wyjściem jest zbiór dokumentów składowych. Operator pozwala na przetwarzanie każdego z dokumentów składowych. Na potrzeby zadania to przetwarzanie jest niekonieczne. Do poprawnej pracy operatora konieczne jest jego odpowiednie skonfigurowanie, gdzie jako *query type* należy wybrać XPath – zaawansowana biblioteka do przetwarzania dokumentów typu XML



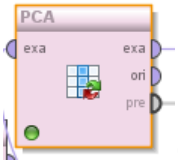
Document to Data – operator dokonuje konwersji zbioru dokumentów na zbiór danych przetwarzanych przez RapidMinera

Operatory związane z analizą danych

Operator służący do klasteryzacji czyli grupowania danych. Jego zadaniem jest znalezienie obiektów które sa podobne do siebie. Uwaga w tym zagadnieniu należy użyć algorytmu *k-Medoids* . Algorytm ten działa podobnie do algorytmu *k-średnich* (k-means) jednakże posiada istotną różnicę, polegającą na tym iż centroidy – wzorce dla każdej z grup są obiektami z oryginalnego zbioru danych

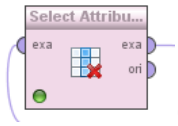


Operator pozwala na wyznaczenie odległości między poszczególnymi obiektami znajdującymi się w dwóch zbiorach danych

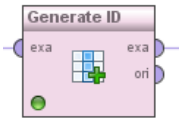


Operator redukcji wymiarowości. Jego zadaniem jest redukcja wymiarowości z przestrzeni wysoko wymiarowej tj. z przypadku w którym mamy dużo kolumn do przestrzeni nisko wymiarowej (mało kolumn) przy czym realizowane jest to tak, aby jak najmniej informacji zostało zgubionej podczas takiej redukcji

Operatory podstawowe



Operator pozwala na wybranie atrybutu/atrybutów (kolumny/kolumn) ze zbioru danych



Operator dodaje nową kolumnę identyfikująca każdy wiersz (example) ze zbioru danych. ID można traktować jako klucz własny tabeli w bazach danych

Operator pozwala odfiltrować pojedyncze wiersze (example) ze zbioru danych (ExampleSet)

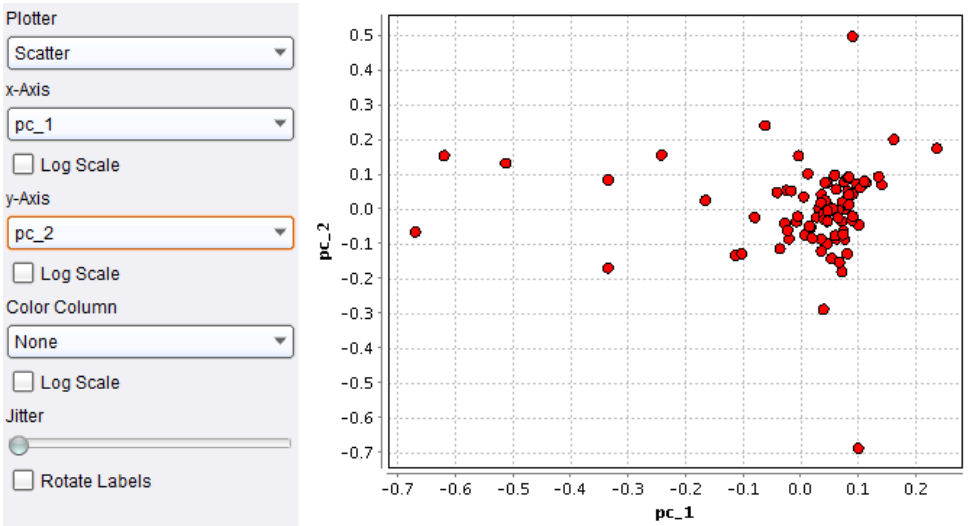
Operator pozwalający na wykonanie operacji *join* z dwóch różnych zbiorów danych RapidMiner'a

Do zrobienia

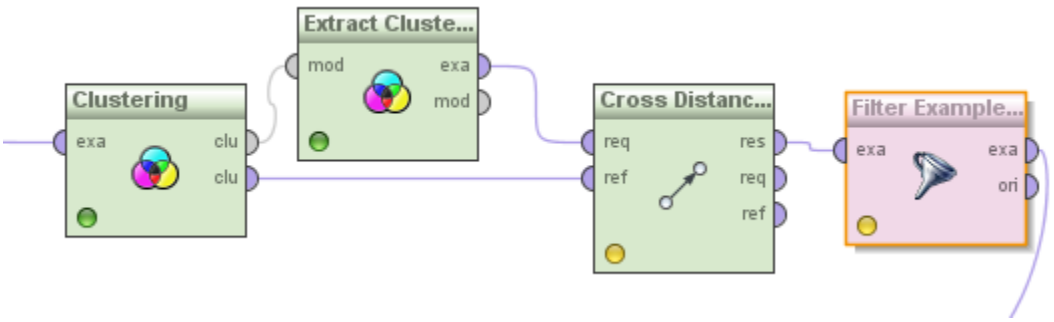
1. Wejdź do Internetu i znajdź forum dyskusyjne z około 30 do 100 wpisów. Wykorzystaj operator *Get page* do pobrania strony do RapidMinera (RM) Zobacz i skomentuj wynik
2. Wykorzystaj operator *Cut documents* do podzielenia strony na obszary odpowiedzialne poszczególnym wpisom. Podczas konfiguracji *xpath queries* w kolumnie *attribute name* wpisz nazwę atrybutu np. *tekst* natomiast w kolumnie *query expression* wykorzystaj polecenie: //h:tag[@class='class_name']/text() Gdzie: tag to nazwa tagu separującego tekst a class_name to nazwa klasy np. dla pojedynczego wpisu o postaci:

```
<div class="post_content">
    &amp;quot;Czy prawdziwy twardy katolik Prof. Chazan wyspowiada&amp;#322; si&amp;#281;
    ju&amp;#380; z romansu i nie&amp;#347;lubnego dziecka z pracownic&amp;#261; szpitala? A
    mo&amp;#380;e tu te&amp;#380; obowi&amp;#261;zuj&amp;#261; podw&amp;oacute;jne standardy?&amp;quot;
    <br /> <br />To taki z niego &amp;quot;katolik&amp;quot;? To ciekawe, ze spora czesc polskich
    kato-talibow to ludzie, ktorzy sami maja rozne swinstwa <br />i swinstewka na sumieniu,
    i czesto (doslownie) az bije od nich hipokryzja!
</div>
```

   Należy wpisać: //h:div[@class= post_content']/text()
   Zobacz i skomentuj wynik
3. Usuń zbędne atrybuty powstałe przez wykorzystacie operatora *Get page* za pomocą operatora *Select attributes.* Konieczne jest aby pozostał jeden atrybut *tekst*
4. Do poprawnego działania reszty procesu przydatne jest wygenerowanie gluczy własnego dla zbioru danych. W tym celu wykorzystaj operator *Generate ID*
5. Korzystając z wiedzy zdobytej na zajęciach poświęconych analizie tekstów (Text Mining) zamień treść wiadomości na reprezentację *worka słów*
   Uwaga: pamiętaj o zamianie wielkości liter, tokenizacji, filtracji krótkich 1 lub2 wyrazowych słów oraz o konwersji polskich znaków diakrytycznych na odpowiedniki „bez ogonków". Do ostatniego zadania wykorzystaj operator *Replace tokens*, zamieniając kolejno ą->a, ó->o, …
   Sprawdź poprawność wyników
6. Gdyby pojawiły się puste dokumenty usuń je. Można to zrobić korzystając z operatorów *Generate aggregation,* który powinien policzyć sumaryczną czyęstość występowania wyrazów, jeśli jest ona =0 wówczas takie wyrazy należy odfiltrować za pomocą operatora F*ilter examples (attribute value filter)*
7. Usuń wygenerowany atrybut z zadania 6 aby nie zakłócał procesu klasteryzacji
8. Dokonaj redukcji wymiarowości za pomocą operatora PCA
   Zobacz wyniki zadania oraz wykonaj ich wizualizację za pomocą polecenia *plot view* w zakładce wynik. Pokaż wynik w postaci zależności PC_1 / PC_2 czyli dwóch składowych głównych po PCA



9. Na uzyskanych wynika przeprowadź proces grupowania za pomocą operatora *k-Medodids*
   Dokonaj wizualizacji jak w punkcie 8, ale ustaw *Color column* na *Cluster*
10. Za pomocą operatora *Extract Cluster Prototypes* dokonaj dokumentów wzorcowych
11. Aby przypisać do dokumentów wzorcowych treść wiadomości wykorzystaj operator *Cross distace* i odfiltruj wszystkie obiekty które mają dystans = 0 (są to obiekty wzorcowe)



   Sprawdź uzyskane wyniki – odczytaj które z wpisów sa wpisami wzorcowymi (odczytaj ich ID)
12. Dokonaj operacji *JOIN* na odnajezionych ID i oryginalnych tekstach uzyskanych w pkt 4
    W wyniku powinny zostać wyświetlone jedynie wzorcowe wpisy na forum.

Dodatek: Xpath

XPath jest standardem przetwarzania dokumentów XML'owych.

Przydatne funkcje:

Wyrażenie	Opis
nazwawęzła	Wybiera wszystkie węzły z określoną nazwą "*nazwawęzła*"
/	Wybiera począwszy od węzła będącego korzeniem
//	Wybiera węzły z dokumentu, które pasują do wyboru, nieważne gdzie one się znajdują
.	Wybiera obecny węzeł

..	Wybiera rodzica obecnego węzła
@	Wybiera atrybut węzła

Przykłady:

```
<books>
   <description>A list of books useful for people first learning how to build web XML web
applictypeions.</description>
   <book type='new'>
      <title>XQuery</title>
      <author>Priscilla Walmsley</author>
      <description>This book is a highly detailed, through and complete tour of the W3C
Query language.  It covers all the
      key aspects of the language as well as</description>
      <formtype>Trade press</formtype>
      <license>Commercial</license>
      <list-price>49.95</list-price>
   </book>
   <book class='1' type='new'>
      <title>XQuery Examples</title>
      <author>Chris Wallace</author>
      <author>Dan McCreary</author>
      <description>This book provides a variety of XQuery example programs and is
designed to work with the eXist open-source ntypeive XML applictypeion
server.</description>
      <formtype>Wiki-books</formtype>
      <license>Cretypeive Commons Sharealike 3.0 typetribution-Non-
commercial</license>
      <list-price>29.95</list-price>
   </book>
   <book>
      <title lang='eng'>XForms Tutorial and Cookbook</title>
      <author>Dan McCreary</author>
      <description>This book is an excellent guide for anyone thtype is just beginning to
learn the XForms standard.  The book
      is focused on providing the reader with simple, but complete examples of how to
cretypee XForms web applictypeions.</description>
      <formtype>Wikibook</formtype>
      <license>Cretypeive Commons Sharealike 3.0 typetribution-Non-
commercial</license>
      <list-price>29.95</list-price>
   </book>
   <book class='1' type='new'>
      <title>XRX: XForms, Rest and XQuery</title>
      <author>Dan McCreary</author>
      <description>This book is an overview of the key architectural and design
ptypeters.</description>
      <formtype>Wikibook</formtype>
      <license>Cretypeive Commons Sharealike 3.0 typetribution-Non-
commercial</license>
      <list-price>29.95</list-price>
   </book>
</books>
```

Przykład	Wynik	Opis
/books/book[1]	<book at="ala"> <title>XQuery</title> <author>Priscilla Walmsley</author> <description>This book is a highly detailed, through and complete tour of the W3C Query language. It covers all the key aspects of the language as well as</description> <format>Trade press</format> <license>Commercial</license> <list-price>49.95</list-price> </book>	Zwrócił pierwszą książkę która jest pod węzłem książki
/books/book[2]/title	<title>XQuery Examples</title>	Zwrócił tytuł drugiej książki
/books/book[last()]/title/text()	XRX: XForms, Rest and XQuery	Zwrócił sam tekst tytułu ostatniej książki – uwaga na polecenie /text() [zwraca treść wpisu] oraz /last() [zwraca ostatni element na liście]
//book[@type='new']/title/text()	Wynik1: XQuery Wynik2: XQuery Examples Wynik3: XRX: XForms, Rest and XQuery	Polecenie szuka wszystkich książek które mają ustawiony atrybut *type='new'*, a spośród nich odczytuje treść tytułu. Uwaga atrybuty wpisujemy począwszy od @ Na początku pojawia się //, co oznacza że szukamy pod dowolnym węzłem ciągu podwęzłów typu book[…]/…
//book[@class='1' and @type='new']/title/text()	Wynik1: XQuery Examples Wynik2: XRX: XForms, Rest and XQuery	Polecenie powoduje wyszukanie węzłów typu książka, w których ustawione sa dwa atrybuty class='1' oraz type='new'
/books/book[list-price>35.00]/title	<title>XQuery</title>	Spośród książek znajdź książkę/książki, których wartość węzła 'last_price' jest większa od 35 i wyświetl ich tytuł
//title[@lang]	<title lang="eng">XForms Tutorial and Cookbook</title>	Znajdź wszystkie te tytuły które

		posiadają atrybut @lang, nie ważne jaką przyjmuje on wartość
/books/book[position()<3] /license	Wynik1: <license>Commercial</license> Wynik2: <license>Cretypeive Commons Sharealike 3.0 typetribution-Non-commercial</license>	Znajduje pierwsze dwie książki spośród książek. Uwaga position() określa pozycję kursora
//book[contains(title, 'XQuery')]/title/text()	Wynik1: XQuery Wynik2: XQuery Examples Wynik3: XRX: XForms, Rest and XQuery	Zwraca treść tytułów tych książek które w polu tytuł zawierają tekst 'XQuery'
//title[contains(text(), 'XQuery')]/../formtype	Wynik1: <formtype>Trade press</formtype> Wynik2: <formtype>Wiki-books</formtype> Wynik3: <formtype>Wikibook</formtype>	Polecenie powoduje odnalezienie wszystkich węzłów w których występuje pole title, które to pole powinno zawierać tekst XQuery, następnie spośród tych węzłów odczytujemy rodzica I z rodzica węzła odczytujemy wartość pola formtype